An Information-theoretic Approach to Privacy

Lalitha Sankar, S. Raj Rajagopalan, and H. Vincent Poor

Abstract— **Ensuring the usefulness of electronic data sources while providing necessary privacy guarantees is an important unsolved problem. This problem drives the need for an overarching analytical framework that can quantify the safety of personally identifiable information (privacy) while still providing a quantifable benefit (utility) to multiple legitimate information consumers. State of the art approaches have predominantly focused on privacy. This paper presents the first information-theoretic approach that promises an analytical model guaranteeing tight bounds of how much utility is possible for a given level of privacy and vice-versa.**

I. The Database Privacy Problem

Information technology and electronic communications have been rapidly applied to almost every sphere of human activity, including commerce, medicine and social networking. The concomitant emergence of myriad large centralized searchable data repositories has made "leakage" of private information such as medical data, credit card information, or social security numbers via data correlation (inadvertently or by malicious design) highly probable and thus an important and urgent societal problem. Unlike the well-studied secrecy problem (e.g., [1]–[3]) in which the protocols or primitives make a sharp distinction between secret and non-secret data, in the *privacy* problem, disclosing data provides informational utility while enabling possible loss of privacy at the same time. In fact, in the course of a legitimate transaction, a user learns some public information, which is allowed and needs to be supported for the transaction to be meaningful, but at the same time he can also learn/infer private information, which needs to be prevented. Thus every user is (potentially) also an adversary. This drives the need for a unified analytical framework that can tell us unequivocally and precisely how safe private data can be (privacy) and simultaneously provide measurable benefit (utility) to multiple legitimate information consumers.

It has been noted that utility and privacy are competing goals: *perfect privacy can be achieved by publishing nothing at all, but this has no utility; perfect utility can be obtained by publishing the data exactly as received, but this offers no privacy* [4]. Utility of a data source is potentially (but not necessarily) degraded when it is restricted or modified to uphold privacy requirements. The central problem of this paper is a precise quantification of the tradeoff between the privacy needs of the *respondents* (individuals represented by the data) and the utility of the *sanitized* (published) data for any data source.

Though the problem of privacy and information leakage has been studied for several decades by multiple research communities (e.g., [4]–[8] and the references therein), the proposed solutions have been both heuristic and application-specific. The recent groundbreaking theory of ϵ-differential privacy [9] from the theoretical computer science community provides the first universal metric of privacy that applies to any numerical database. We seek to address the open question of a universal and analytical characterization that provides a tight privacy-utility tradeoff using tools and techniques from information theory.

Rate distortion theory is a natural choice to study the utility-privacy tradeoff; utility can be quantified via fidelity which, in turn, is related to *distortion*, and privacy can be quantified via *equivocation*. Our key insight is captured in the following theorem which is presented in this paper: for a data source with private and public data, *minimizing the information disclosure rate* sufficiently to satisfy the desired utility for the public data *is equivalent to maximizing the privacy* for the private data. In a sparsely referenced paper [10] from three decades ago, Yamamoto developed the tradeoff between rate, distortion, and equivocation for a specific and simple source model. In this paper, we show via the above summarized theorem that Yamamoto's formalism can be translated into the language of data disclosure. Furthermore, we develop a framework that allows us to model data sources, specifically databases, develop application independent utility and privacy metrics, quantify the fundamental bounds on the utility-privacy tradeoffs, and develop a side-information model for dealing with questions of external knowledge.

The paper is organized as follows. We present channel model and preliminaries in Section II. The main result and the proof are developed in Section III. We discuss the results and present numerical examples in Section IV. We conclude in Section V.

II. The Database Privacy Problem

A. Problem Definition

While the problem of quantifying the utility/privacy problem applies to all types of data sources, we start our study with databases because they are highly structured and historically better studied than other types of sources. A database is a table (matrix) whose rows represent individual entries and whose columns represent the *attributes* of each entry [5]. For example, the attributes of each entry in a healthcare database typically include name, address, social

This research is supported in part by the U. S. National Science Foundation under Grant CCF-1016671, the Air Force Office of Scientific Research under Grant FA9550-09-1-0643, and by a fellowship from the Princeton University Council on Science and Technology.

L. Sankar and H. V. Poor are with the Department of Electrical Engineering, Princeton University, Princeton, NJ, USA. S. Raj Rajagopalan is with HP Labs, Princeton, NJ, USA. {lalitha,poor@princeton.edu,raj.rajagopalan@hp.com}

security number (SSN), gender, and a collection of medical information, and each entry contains the information pertaining to an individual. Messages from a *user* to a database are called *queries* and, in general, result in some numeric or non-numeric information from the database termed the *response*.

The goal of privacy protection is to ensure that, to the extent possible, the user's knowledge is not increased beyond strict predefined limits by interacting with the database. The goal of utility provision is, generally, to maximize the amount of information that the user can receive. Depending on the relationships between attributes, and the distribution of the actual data, a response may contain information that can be inferred beyond what is explicitly included in the response. The privacy policy defines the information that should not be revealed explicitly or by inference to the user and depends on the context and the application. For example, in a database on health statistics, attributes such as name and SSN may be considered private data, whereas in a state motor vehicles database only the SSN is considered private. The challenge for privacy protection is to design databases such that responses do not reveal information contravening the privacy policy.

B. Current Approaches and Metrics

The problem of privacy in databases has a long and rich history stretching back to the 1970s and space restrictions preclude any attempt to do full justice to the different approaches that have been considered along the way. While there have been many heuristic approaches to privacy, we only present the major milestones in privacy research on creating quantitative privacy metrics. Since privacy is a requirement that appears in many diverse contexts, a robust and formal notion of privacy that satisifies most, if not all, requirements is a tricky proposition and there have been many attempts at a definition. The reader is referred to the excellent survey by Dwork [11] for a detailed history of the field. The problem of privacy was first exposed by census statisticians who were required to publish statistics related to census functions but without revealing any particulars of individuals in the census databases. An early work by Dalenius [6] reveals the depth to which this problem was considered. Several early attempts were made to publish census data using ad hoc techniques such as sub-sampling. However, the first widely reported attempt at a formal definition of privacy was by Sweeney [7]. The concept of *k-anonymity* proposed by Sweeney captures the intuitive notion of privacy that every individual entry should be indistinguishable from $(k-1)$ other entries for some large value of k. This notion of anonymity for database respondents is analogous to similar proposals that were made for anonymity on the Internet such as crowds [12]. More recently, researchers in the data mining community have proposed to quantify the privacy loss resulting from data disclosure as the mutual information between attribute values in the original and perturbed data sets, both modeled as random variables [8].

The approaches considered in the literature have centered on the correct application of *perturbation* (also called *sanitization*), which encompasses a general class of database modification techniques that ensure that a user interacts only with a modified database that is derived from the original (e.g.: [4], [6]–[8]). Most of the these perturbation approaches, with the exception of differential privacy-based ones, are heuristic and application-specific and often focus on additive noise approaches.

Differential privacy: More recently, privacy approaches for statistical databases has been driven by the differential privacy definition [9], [13]–[15]. In these papers, the authors take the view that privacy of an individual in a database is related to the ability of an adversary to detect whether that individual's data is in that database or not. Motivated by cryptographic models, they formalize this intuition by defining the difference in the adversary's outputs when presented with two databases D and $D^{'}$ that are identical except in one row.

Definition 1 ([11]): A function $\mathcal{K}$ gives ϵ-differential privacy if for all databases $D, D^{'}$ defined as above, and all $S \subseteq$ Range($\mathcal{K}$),

$$\Pr[\mathcal{K}(D) \in S] \leq exp(\epsilon) \cdot \Pr\left[\mathcal{K}\left(D^{'}\right) \in S\right] \qquad (1)$$

where the probability space in each case is over the coin flips of $\mathcal{K}$.

It is important to make two observations regarding the above definition. First, the probabilities in definition 1 are over the actions of the function $\mathcal{K}$ and not over the distribution of D; in other words, the definition is independent of the distribution from which D may be sampled. Second, Definition 1 guarantees that the presence or absence of an individual row in the database makes very little difference to the output of the adversary as required, and thus, provides a precise privacy guarantee to any individual in the database.

More recently, Dwork *et al.* [15] also provide a mechanism for achieving ϵ-differential privacy universally for statistical queries (queries that map subsets of database entries to real numbers) which we summarize below. Let $Z \sim Lap(b)$ represent a Laplacian distributed random variable with parameter b. If $b = 1/\epsilon$ we have that the density at z is proportional to $\exp(-4|z|)$ and for any (z, z') such that $|z - z'| \leq 1$, $\Pr(z)$ and $\Pr(z^{'})$ are within a factor of e^{ϵ}. The following proposition shows that it is possible to achieve ϵ-differential privacy for a given statistical query class for suitable choice of the Laplacian parameter.

Proposition 1 ([15]): For any statistical query $f : D \to \mathcal{R}$, the mechanism L that adds independently generated noise to the output terms with distribution $Lap(\,\Delta_f/\,\epsilon)$ guarantees ϵ-differential privacy where $\Delta_f = \max\left|f(D) - f(D^{'})\right|$ for D, $D^{'}$ which are different in exactly one row.

Proposition 1 is the most significant milestone in the theory of privacy because it provides a method to guarantee a strong but quantifiable notion of privacy for statistical databases independent of their content. Furthermore, the noise distribution can be chosen after seeing the query, so that the noise level can be adjusted adaptively when presented with a sequence of queries. However, one constraint in using Proposition 1 to define ϵ is that Δf may be difficult to estimate – a loose bound on Δf may result in an overly large ϵ, thereby resulting in a possible degradation of utility.

To date, privacy has been the main focus of most work in this area. Indeed, Dwork [9] says explicitly that privacy is paramount in their work. However, databases exist to be useful and implementing sanitization techniques may hurt the usefulness of the database while safeguarding privacy. In much of the earlier work on database privacy, the utility is implicit. For exmple, Sweeney assumes that the databases can be k-anonymized and still maintain usefulness. However, without a relationship between k and some formal notion of usefulness, it is impossible to say what a reasonable value of k should be in reality. Similarly, utility in privacy-preserving techniques such as clustering [4] and histograms [16] is assumed to be guaranteed as a direct result of the methods used; for example, in [16] it is shown that approximation algorithms that can run on original histograms can also run on the sanitized histograms with a degradation of performance. Clustering, a common sanitization technique [4], [17], [18], is claimed to maintain utility as a result of the following property: all points in a cluster are mapped to the cluster center, so no point is moved more than the diameter of the largest cluster.

The differential privacy model uses additive noise for sanitization which in turn suggests a utility metric related to the accuracy of the sanitized database. The Laplacian noise model was chosen for achieving differential privacy in part because the mean and mode are zero, in which case no noise is added in most cases. The privacy parameter ϵ is inversely related to the variance of the added noise – a better privacy guarantee requires a smaller ϵ which in turn implies higher variance. The accuracy of a sanitized database as a whole is inversely related to the privacy requirement. Determining the appropriate range of ϵ so that both privacy and accuracy requirements are balanced requires knowledge of the specific application. As an example, in the case of learning, recent results [19] in the area of *private learning* bound the extent to which the performance (i.e. accuracy) of certain kinds of classifers degrade when the training data is sanitized using the L mechanism in Proposition 1. In such cases, it is possible to have both, differential privacy with a known ϵ, as well as quantified utility loss for the application under consideration.

C. Privacy vs. Secrecy

It is important to contrast the privacy problem from the well-studied (cryptographic and information-theoretic) *secrecy* problem where the task is to stop specific information from being received by untrusted third parties (eavesdroppers, wire-tappers, and other kinds of adversaries). In the *private information retrieval* model [20], the privacy problem is inverted in that the adversary is the database from whom the user wants to keep his *query* secret. In the secure multi-party computation model [21], each player wishes to keep his *entire input* secret from the other players while jointly computing a function on all the inputs. In all these problems, a specific data item is clearly either secret or public, whereas in the privacy problem, the same data while providing informational utility to the user can reveal private information about the individuals represented by the data. This eliminates the possibility of using secrecy techniques such as a specific model of the adversary or of harnessing any computing [22] or physical advantages such as secret keys, channel differences, or side information [23].

III. An Information-Theoretic Approach

A. Model for Databases

Circumventing the semantic issue: In general, utility and privacy metrics tend to be application specific. Focusing our efforts on developing an analytical model, we propose to capture a canonical database model and representative abstract metrics. Such a model will circumvent the classic privacy issues related to the semantics of the data by assuming that there exist forward and reverse maps of the data set to the proposed abstract format (for e.g., a string of bits or a sequence of real values). Such mappings are often implicitly assumed in the privacy literature [4], [8], [9]; our motivation for making it explicit is to separate the semantic issues from the abstraction and apply Shannon-theoretic techniques.

Model: Our proposed model focuses on large databases with K attributes per entry. Let $X_k \in \mathcal{X}_k$ be a random variable denoting the k^{th} attribute, $k = 1, 2, \ldots, K$, and let $\mathbf{X} \equiv (X_1, X_2, \ldots, X_K)$. A database d with n rows is a sequence of n independent observations of $\mathbf{X}$ from the distribution

$$p_{\mathbf{X}}(\mathbf{x}) = p_{X_1 X_2 \ldots X_K}(x_1, x_2, \ldots, x_K) \qquad (2)$$

which is assumed to be known to both the designers and users of the database. Our simplifying assumption of row independence holds generally (but not always) as correlation is typically across attributes and not across entries. We write $\mathbf{X}^n = (X_1^n, X_2^n, \ldots, X_K^n)$ to denote the n independent observations of $\mathbf{X}$. This database model is universal in the sense that most practical databases can be mapped to this model.

A joint distribution in (2) models the fact that the attributes in general are correlated and can reveal information about one another. In addition to the revealed information, a user of a database can have access to correlated side information from other information sources. We model the side-information as an n-length sequence Z^n which is correlated with the database entries via a joint distribution $p_{\mathbf{X}Z}(\mathbf{x}, z)$.

Public and private variables: We consider a general model in which some attributes need to be kept private while the source can reveal a function of some or all of the attributes. We write $\mathcal{K}_r$ and $\mathcal{K}_h$ to denote sets of private (subscript h for hidden) and public (subscript r for revealed) attributes, respectively, such that $\mathcal{K}_r \cup \mathcal{K}_h = \mathcal{K} \equiv \{1, 2, \ldots, K\}$. We further denote the corresponding collections of public and private attributes by $\mathbf{X}_r \equiv \{X_k\}_{k \in \mathcal{K}_r}$ and $\mathbf{X}_h \equiv \{X_k\}_{k \in \mathcal{K}_h}$, respectively. Our notation allows for an attribute to be both public and private; this is to account for the fact that a database may need to reveal a function of an attribute while keeping the attribute itself private. In general, a database can choose to keep public (or private) one or more attributes ($K > 1$). Irrespective of the number of private attributes, a non-zero utility results only when the database reveals an appropriate function of some or all of its attributes.

Special cases: For $K = 1$, the lone attribute of each entry (row) is both public and private, and thus, we have $X \equiv X_r \equiv X_h$. Such a model is appropriate for data mining [8]; for a more general case in which $K_h = K_r = K$, we obtain a model for census [4], [6] data sets in which utility generally is achieved by revealing a function of every entry of the database while simultaneously ensuring that no entry is perfectly revealed. For $K = 2$ and $\mathcal{K}_h \cup \mathcal{K}_r = \mathcal{K}$ and $\mathcal{K}_h \cap \mathcal{K}_r = \emptyset$, we obtain the Yamamoto model in [10].

B. Metrics: The Privacy and Utility Principle

Even though utility and privacy measures tend to be specific to the application, there is a fundamental principle that unifies all these measures in the abstract domain. The aim of a privacy-preserving database is to provide some measure of utility to the user while at the same time guaranteeing a measure of privacy for the entries in the database.

A user perceives the utility of a perturbed database to be high as long as the response is similar to the response of the unperturbed database; thus, the utility is highest of an unperturbed database and goes to zero when the perturbed database is completely unrelated to the original database. Accordingly, our utility metric is an appropriately chosen average ‘distance’ function between the original and the perturbed databases. Privacy, on the other hand, is maximized when the perturbed response is completely independent of the data. Our privacy metric measures the difficulty of extracting any private information from the response, i.e., the amount of uncertainty or *equivocation* about the private attributes given the response.

C. Utility-Privacy Tradeoffs

1) A Privacy-Utility Tradeoff Model: We now propose a privacy-utility model for databases. *Our primary contribution is demonstrating the equivalence between the database privacy problem and a source coding problem with additional privacy constraints*. A primary motivation for our approach is the observation that database sanitization is traditionally the process of distorting the data to achieve some measure of privacy. For our abstract universal database model, sanitization is thus a problem of mapping a set of database entries to a different set subject to specific utility and privacy requirements.

Our notation below relies on this abstraction. Recall that a database d with n rows is an instantiation of $\mathbf{X}^n$. Thus, we will henceforth refer to a real database d as an *input sequence* and to the corresponding sanitized database (SDB) d' as an *output sequence*. When the user has access to side information, the *reconstructed sequence* at the user will in general be different from the SDB sequence.

Our coding scheme consists of an encoder F_E which is a mapping from the set of all input sequences (i.e., all databases d chosen from an underlying distribution) to a set of indices $\mathcal{W} \equiv \{1, 2, \ldots, M\}$ and an associated table of output sequences (each of which is a d') with a one-to-one mapping to the set of indices given by

$$F_E : (\mathcal{X}_1^n \times \mathcal{X}_2^n \times \ldots \times \mathcal{X}_k^n)_{k\in\mathcal{K}_{enc}} \to \mathcal{W} \equiv \{SDB_k\}_{k=1}^M \tag{3}$$

where $\mathcal{K}_r \subseteq \mathcal{K}_{enc} \subseteq \mathcal{K}$ and $M = 2^{nR}$ is the number of output (sanitized) sequences created from the set of all input sequences. The encoding rate R is the number of bits per row (without loss of generality, we assume n rows in d and d') of the sanitized database. The encoding F_E in (3) includes both public and private attributes in order to model the general case in which the sanitization depends on a subset of all attributes.

A user with a view of the SDB (i.e., an index $w \in \mathcal{W}$ for every d) and with access to side information Z^n, whose entries Z_i, $i = 1, 2, \ldots, n$, take values in the alphabet $\mathcal{Z}$, reconstructs the database d' via the mapping

$$F_D : \mathcal{W} \times \mathcal{Z}^n \to \left\{\hat{\mathbf{x}}_{r,m}^n\right\}_{m=1}^M \in \left(\prod\nolimits_{k\in\mathcal{K}_r} \hat{\mathcal{X}}_k^n\right) \tag{4}$$

where $\hat{\mathbf{X}}_r^n = F_D\left(F_E\left(\mathbf{X}^n\right)\right)$.

A database may need to satisfy multiple utility constraints for different (disjoint) subsets of attributes, and thus, we consider a general framework with $L \geq 1$ utility functions that need to be satisfied. Relying on the distance based utility principle, we model the l^{th} utility, $l = 1, 2, \ldots, L$, via the requirement that the average *distortion* Δ_l of the revealed variables is upper bounded, for some $\epsilon > 0$, as

$$u_l : \Delta_l \equiv \mathbb{E}\left[\frac{1}{n}\sum\nolimits_{i=1}^n g\left(\mathbf{X}_{r,i}, \hat{\mathbf{X}}_{r,i}\right)\right] \leq D_l + \epsilon, \qquad l = 1, 2, \ldots, L, \tag{5}$$

where $g\left(\cdot,\cdot\right)$ denotes a distortion function, $\mathbb{E}$ is the expectation over the joint distribution of $(\mathbf{X}_r, \hat{\mathbf{X}}_r)$, and the subscript i in $\mathbf{X}_{r,i}$ and $\hat{\mathbf{X}}_{r,i}$ denotes the i^{th} entry of $\mathbf{X}_r^n$ and $\hat{\mathbf{X}}_r^n$, respectively. Examples of distance-based distortion functions include the Euclidean distance for Gaussian distributed database entries, the Hamming distance for binary input and output sequences, and the Kullback-Leibler (K-L) ‘distance’ comparing the input and output distributions.

Having argued that a quantifiable uncertainty captures the underlying privacy principle of a database, we model the uncertainty or equivocation about the private variables using the entropy function as

$$p : \Delta_p \equiv \frac{1}{n} H\left(\mathbf{X}_h^n | W, Z^n\right) \geq E - \epsilon, \tag{6}$$

i.e., we require the average number of uncertain bits per entry to be lower bounded by E. The case in which side information is not available at the user is obtained by simply setting $Z^n = 0$ in (4) and (6).

The utility and privacy metrics in (5) and (6), respectively, capture two aspects of our universal model: a) both represent averages by computing the metrics across all database instantiations d, and b) the metrics bound the average distortion and privacy per entry. Thus, as the likelihood of the non-typical sequences decreases exponentially with increasing n (very large databases), these guarantees apply nearly uniformly to all (typical) entries. Our general model also encompasses the fact that the exact mapping from the distortion and equivocation domains to the utility and privacy domains, respectively, can depend on the application domain. We write $D \equiv (D_1, D_2, \ldots, D_L)$ and $\Delta \equiv (\Delta_1, \Delta_2, \ldots, \Delta_L)$. Based on our notation thus far, we define the utility-privacy tradeoff region as follows.

Definition 2: The utility-privacy tradeoff region $\mathcal{T}$ is the set of all feasible utility-privacy tuples (D, E) for which there exists a coding scheme (F_E, F_D) given by (3) and (4), respectively, with parameters (n, M, Δ, Δ_p) satisfying the constraints in (5) and (6).

2) Equivalence of Utility-Privacy and Rate-Distortion-Equivocation: We now present an argument for the equivalence of the above utility-privacy tradeoff analysis with a rate-distortion-equivocation analysis of the same source. For the database source model described here, a classic lossy source coding problem is defined as follows.

Definition 3: The set of tuples (R, D) is said to be feasible (achievable) if there exists a coding scheme given by (3) and (4) with parameters (n, M, Δ) satisfying the constraints in (5) and a rate constraint

$$M \leq 2^{n(R+\epsilon)}. \quad (7)$$

When an additional privacy constraint in (6) is included, the source coding problem becomes one of determining the achievable rate-distortion-equivocation region defined as follows.

Definition 4: The rate-distortion-equivocation region $\mathcal{R}$ is the set of all tuples (R, D, E) for which there exists a coding scheme given by (3) and (4) with parameters (n, M, Δ, Δ_p) satisfying the constraints in (5), (6), and (7). The set of all feasible distortion-equivocation tuples (D, E) is denoted by $\mathcal{R}_{D-E}$, the equivocation-distortion function in the D-E plane is denoted by $\Gamma(D)$, and the distortion-equivocation function which quantifies the rate as a function of both D and E is denoted by $R(D, E)$.

Thus, a rate-distortion-equivocation code is by definition a (lossy) source code satisfying a set of distortion constraints that achieves a specific privacy level for every choice of the distortion tuple. In the following theorem, we present a basic result capturing the precise relationship between $\mathcal{T}$ and $\mathcal{R}$. To the best of our knowledge, this is the first analytical result that quantifies a tight relationship between utility and privacy. We briefly sketch the proof here; details can be found in [24].

Theorem 1: For a database with a set of utility and privacy metrics, the tightest utility-privacy tradeoff region $\mathcal{T}$ is the distortion-equivocation region $\mathcal{R}_{D-E}$.

Proof: The crux of our argument is the fact that for any feasible utility level D, choosing the minimum rate $R(D, E)$, ensures that the least amount of *information* is revealed about the source via the reconstructed variables. This in turn ensures that the maximum privacy of the private attributes is achieved for that utility since, in general, the public and private variables are correlated. For the same set of utility constraints, since such a rate requirement is not a part of the utility-privacy model, the resulting privacy achieved is at most as large as that in $\mathcal{R}_{D-E}$ (see Fig. 1(a)). ■

Implicit in the above argument is the fact that a utility-privacy achieving code does not perform any better than a rate-distortion-equivocation code in terms of achieving a lower rate (given by $\log_2 M/n$) for the same distortion and privacy constraints. We can show this by arguing that if such a code exists then we can always find an equivalent source coding problem for which the code would violate Shannon's source coding theorem [25]. An immediate consequence of this is that a distortion-constrained source code suffices to preserve a desired level of privacy; in other words, *the utility constraints require revealing data which in turn comes at a certain privacy cost that must be borne and vice-versa*. We capture this observation in Fig. 1(b) where we contrast existing privacy-exclusive and utility-exclusive regimes (extreme points of the utility-privacy tradeoff curve) with our more general approach of determining the set of feasible utility-privacy tradeoff points.

From an information-theoretic perspective, the power of Theorem 1 is that it allows us to study the larger problem of database utility-privacy tradeoffs in terms of a relatively familiar problem of source coding with privacy constraints. As noted previously, this problem has been studied for a specific source model by Yamamoto and here we expand his elegant analysis to arbitrary database models including those with side information at the user. Rate for the database can be interpreted as the number of revealed information bits (precision) per row. Our result shows the tight relationship between utility, privacy, and precision – fixing the value of any one determines the other two; for example, fixing the utility (distortion D) precisely quantifies the maximal privacy $\Gamma(D)$ and the minimal precision $R(D, E)$ for any E bounded by $\Gamma(D)$.

3) Capturing the Effects of Side-Information: It has been illustrated that when a user has access to an external data source (which is not part of the database under consideration) the level of privacy that can be guaranteed changes [7], [9]. We cast this problem in information-theoretic terms as a side information problem.

In an extended version [24] of this work, we develop the tightest utility-privacy tradeoff region for the three cases of a) no side information ($L = 1$ case studied in [10]), b) side information only at the user, and c) side information at both the source (database) and the user. We present a result for the case with side information at the user only and for simplicity, we assume a single utility function, i.e., $L = 1$. The proof uses an auxiliary random variable U along the lines of source coding with side information [26] and bounds the equivocation just as in [10, Appendix 1]. The following theorem defines the bounds on the region $\mathcal{R}$ in Definition 4 via the functions $\Gamma(D)$ and $R(D, E)$ where

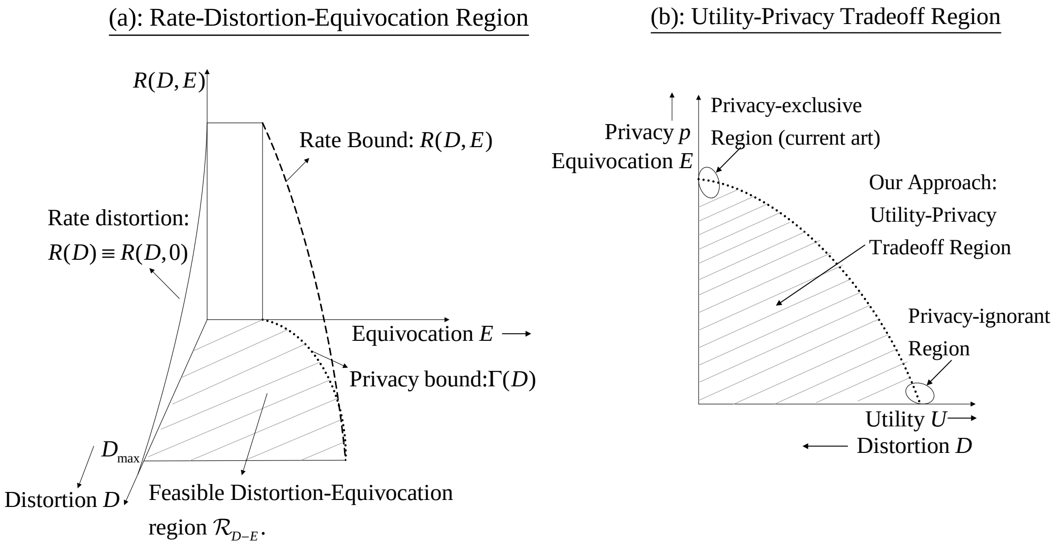

Fig. 1. (a) Rate Distortion Equivocation Region; (b) Utility-Privacy Tradeoff Region.

$\Gamma(D)$ bounds the maximal achievable privacy and $R(D,E)$ is the minimal information rate (see Fig. 1(a)) for very large databases $(n \to \infty)$. The proof follows along the lines of Yamamoto's proof in [10, Appendix 1] and is skipped in the interest of space.

Theorem 2: For a database with side information available only at the user, the functions $\Gamma(D)$ and $R(D,E)$ and the regions $\mathcal{R}_{D-E}$ and $\mathcal{R}$ are given by

$$\Gamma(D) = \sup_{p(\mathbf{x}_r,\mathbf{x}_h)p(u|\mathbf{x}_r,\mathbf{x}_h)\in\mathcal{P}(D)} H(\mathbf{X}_h|UZ) \tag{8}$$

$$R(D,E) = \inf_{p(\mathbf{x}_r,\mathbf{x}_h)p(u|\mathbf{x}_r,\mathbf{x}_h)\in\mathcal{P}(D,E)} I(\mathbf{X}_h\mathbf{X}_r;U) - I(Z;U) \tag{9}$$

$$\mathcal{R}_{D-E} = \{(D,E) : D \ge 0, 0 \le E \le \Gamma(D)\} \tag{10}$$

$$\mathcal{R} = \{(R,D,E) : D \ge 0, 0 \le E \le \Gamma(D), R \ge R(D,E)\} \tag{11}$$

where $\mathcal{P}(D,E)$ is the set of all $p(\mathbf{x}_r,\mathbf{x}_h,z)p(u|\mathbf{x}_r,\mathbf{x}_h)$ such that $\mathbb{E}\left[d\left(\mathbf{X}_r, g(U,Z)\right)\right] \le D$ and $H(\mathbf{X}_h|UZ) \ge E$, while $\mathcal{P}(D)$ is defined as

$$\mathcal{P}(D) \equiv \bigcup\nolimits_{H(\mathbf{X}_h|\mathbf{X}_r Z)\le E\le H(\mathbf{X}_h|Z)} \mathcal{P}(D,E). \tag{12}$$

While Theorem 2 applies to a variety of database models, it is extremely useful in quantifying the utility-privacy tradeoff for the following special cases of interest.

i) *The single database problem* (i.e., no side information): *SDB is revealed*. Here, we have $Z = 0$ and $U = \hat{X}_r$, i.e., the reconstructed vectors seen by the user are the same as the SDB vectors.

ii) *Completely hidden private variables*: *Privacy is completely a function of the statistical relationship between public, private, and side information data*. The expression for $R(D,E)$ in (9) assumes the most general model of encoding both the private and the public variables. When the private variables can only be deduced from the revealed variables, i.e., $\mathbf{X}_h - \mathbf{X}_r - U$ is a Markov chain, the expression for $R(D,E)$ in (9) will simplify to the Wyner-Ziv source coding formulation [26], thus clearly demonstrating that the privacy of the hidden variables is a function of both the correlation between the hidden and revealed variables and the distortion constraint.

iii) *Census and data mining problems without side information*: *Information rate completely determines the degree of privacy achievable*. For $Z = 0$, setting $\mathbf{X}_r = \mathbf{X}_h \equiv \mathbf{X}$ (such that $U = \hat{\mathbf{X}}$), we obtain the census/data mining problem discussed earlier. In general, due to an additional equivocation constraint, $R(D,E) \ge R(D)$; however, for this case in which all the attributes in the database are public, since $\Gamma(D) = H(\mathbf{X}) - R(D,E) \le H(\mathbf{X}) - R(D)$, and $R(D)$ is achievable using a rate-distortion code, the largest possible equivocation is also achievable. Our analysis thus formalizes the intuition in [8] for using the mutual information as an estimate of the privacy lost. However in contrast to [8] in which the underlying perturbation model is an additive noise model, we assume a perturbation model most appropriate for the input statistics, i.e., the stochastic relationship between the output and input variables is chosen to minimize the rate of information transfer.

IV. Illustration of Results

We illustrate our results for two types of databases: one, a *categorical* database and the other a *numerical* database. Categorical data are typically discrete data sets comprising of information such as gender, social security numbers and zipcodes that provide (meaningful) utility only if they are mapped within their own set. On the other hand, without loss of generality numeric data can be assumed to belong to the set of real numbers. In general, a database will have a mixture of categorical and numerical attributes but for the purpose of illustration, we assume that the database is of one type or the other, i.e., every attribute is of the same kind. In both cases, we assume a single utility (distortion) function. We discuss each example in detail below.

Example 1: Consider a categorical database with $K \geq 1$ attributes. In general, the k^{th} attribute X_k takes values in a discrete set $\mathcal{X}_k$ of cardinality M_k. For our example, we model the utility as a single distortion function of all attributes, and therefore, it suffices to view each entry (a row of all K attributes) of the database as generated from a single source X of cardinality M, i.e., $X \sim p(x)$, $x \in \{1, 2, \ldots, M\}$. For this arbitrary discrete source model, we assume that the output sample space $\hat{\mathcal{X}} = \mathcal{X}$ and consider the generalized Hamming distortion as the utility function such that the average distortion D is given by

$$D = E\left[d(X, \hat{X})\right] = \Pr\left\{X \neq \hat{X}\right\}. \tag{13}$$

For $K = 1$, one can show that $R(D, E) \equiv R(D)$ [24]; this is because the maximum achievable equivocation is bounded as $\Gamma(D) = H(X) - R(D, E) \leq H(X) - R(D)$ with equality when $R(D)$ is achievable. It has been shown by Erokhin [27] and Pinkston [28] that $R(D)$ is achieved by upside down waterfilling such that

$$p(\hat{x}) = \frac{(p(x) - \lambda)^+}{\sum_{x \in \mathcal{X}} (p(x) - \lambda)^+} \tag{14}$$

and the 'test channel' is given by

$$p(x|\hat{x}) = \begin{cases} \overline{D}, & x = \hat{x} \\ \lambda, & x \neq \hat{x}, x \in \hat{\mathcal{X}}_{\text{supp}} \\ p_k, & x = k \notin \hat{\mathcal{X}}_{\text{supp}} \end{cases} \tag{15}$$

where $\overline{D} = 1 - D$, λ is chosen such that $\sum_{\hat{x}} p(\hat{x}) p(x|\hat{x}) = p(x)$, $p_k = p\,(x = k)$, and $\hat{\mathcal{X}}_{\text{supp}} = \{x : p(x) - \lambda > 0\}$. The maximum achievable equivocation, and hence, the largest utility-privacy tradeoff region is

$$\Gamma(D) = -\overline{D} \log \overline{D} - \left|\hat{\mathcal{X}}_{\text{supp}}\right| \lambda \log \lambda - \sum_{k \notin \hat{\mathcal{X}}_{\text{supp}}} p_k \log p_k. \tag{16}$$

Remark 1: The distortion function chosen in (13) captures the fact that for categorical data the utility (fidelity) of the revealed data is reduced if any entry is changed from its original value. The optimal upside down waterfilling solution in (14) has the effect of 'flattening' the output distribution, and thus, as in (14) the source samples with very high or very low probabilities (relative to the waterfilling level) are ignored (thereby minimizing the information transfer rate). This in turn maximizes the privacy achieved since the outliers that are easiest to infer are eliminated. Eliminating outliers, referred to as information suppression or aggregation, is the privacy-preserving technique of choice for the statistics community .

Example 2: In this example we model a numerical database. We consider a $K = 2$ database where both attributes X and Y are jointly Gaussian with zero means and variances σ_X^2 and σ_Y^2, respectively, and with correlation coefficient $\rho = E\left[XY\right] / \left(\sigma_X \sigma_Y\right)$. This model applies for numeric data such as height and weight measures which are generally assumed to be normally distributed. We assume that for every entry only one of the two attributes, say X, is revealed while the other, say Y, is hidden such that $Y - X - \hat{X}$ forms a Markov chain. The rate-distortion-equivocation region for this case can be obtained directly from Yamamoto's results [10] with appropriate substitution for a jointly Gaussian source. Furthermore, due to the Markov relationship between of X, Y, and $\hat{X}$, the minimization of $I(X; \hat{X})$ is strictly over $p(\hat{x}|x)$, and thus, simplifies to the familiar rate-distortion problem for a Gaussian source X which in turn is achieved by choosing the reverse channel from $\hat{X}$ to X as an additive white Gaussian noise channel with variance D (average distortion). The maximal equivocation achieved thus is

$$\Gamma(D) = \sigma_Y^2 \left[\left(1 - \rho^2\right) + \rho^2 D \left/\sigma_X^2\right.\right], \quad D \leq \sigma_X^2. \tag{17}$$

Therefore, $\Gamma(D)$ is a minimum for $D = 0$ (X revealed perfectly) in which case only the data independent of X in Y can be private, and is a maximum equal to the entropy of Y at the maximum distortion $D = \sigma_X^2$. Thus, the largest utility-privacy tradeoff region is simply the region enclosed by $\Gamma(D)$.

V. Concluding Remarks

We have presented an abstract model for databases with an arbitrary number of public and private variables, developed application-independent privacy and utility metrics, and used rate distortion theory to determine the fundamental utility-privacy tradeoff limits. Future work includes eliminating the row independence (i.i.d) assumption, modeling and studying tradeoffs for multiple query databases, and relating current approaches in computer science and our universal approach. An equally pertinent question is to understand whether our formalism can be extended to study privacy-utility tradeoffs for less structured datasets as well as social networks.

References

[1] R. Rivest, A. Shamir, and L. Adleman, "A method for obtaining digital signatures and public-key cryptosystem," *Communications of the ACM*, vol. 21, no. 2, pp. 120–126, Feb. 1978.

[2] I. Csiszár and P. Narayan, "Secrecy capacities for multiple terminals," *IEEE Trans. Inform. Theory*, vol. 50, no. 12, pp. 3047–3061, Dec. 2004.

[3] A. D. Wyner, "The wire-tap channel," *Bell Sys. Tech. J.*, vol. 54, pp. 1355–1387, 1975.

[4] S. Chawla, C. Dwork, F. McSherry, A. Smith, and H. Wee, "Towards privacy in public databases," in *Proc. 2nd IACR Theory Crypto. Conf.*, Cambridge, MA, Feb. 2005, pp. 363–385.

[5] N. R. Adam and J. C. Wortmann, "Security-control methods for statistical databases: A comparative study," *ACM Computing Surveys*, vol. 21, no. 4, pp. 515–556, 1989.

[6] T. Dalenius, "Finding a needle in a haystack - or identifying anonymous census records," *J. Official Stats.*, vol. 2, no. 3, pp. 329–336, 1986.

[7] L. Sweeney, "*k*-anonymity: A model for protecting privacy," *IEEE Trans. Inform. Theory*, vol. 10, no. 5, pp. 557–570, 2002.

[8] D. Agrawal and C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," in *Proc. 20th Symp. Principles of Database Systems*, Santa Barbara, CA, May 2001.

[9] C. Dwork, "Differential privacy," in *Proc. 33rd Intl. Colloq. Automata, Lang., Prog.*, Venice, Italy, July 2006.

[10] H. Yamamoto, "A source coding problem for sources with additional outputs to keep secret from the receiver or wiretappers," *IEEE Trans. Inform. Theory*, vol. 29, no. 6, pp. 918–923, Nov. 1983.

[11] C. Dwork, "A firm foundation for private data analysis," Jan. 2011, http://research.microsoft.com/apps/pubs/?id=116123, to appear in Communications of the ACM.

[12] M. K. Reiter and A. D. Rubin, “Anonymous web transactions with crowds,” *Communications of the ACM*, vol. 42, no. 2, pp. 32–48, 1999.

[13] C. Dwork and A. Smith, “Differential privacy for statistics: what we know and what we want to learn,” in *Proc. NCHS/CDC Data Confidentiality Workshop*, Hyattsville, MD, May 2008.

[14] C. Dwork, “Differential privacy: A survey of results,” in *Theory and Applications of Models of Computation: Lecture Notes in Computer Science*. New York:Springer, Apr. 2008.

[15] C. Dwork, F. McSherry, K. Nissim, and A. Smith, “Calibrating noise to sensitivity in private data analysis,” in *Proc. 3rd IACR Theory Crypto. Conf.*, New York, NY, Mar. 2006.

[16] S. Chawla, C. Dwork, F. McSherry, and K. Talwar, “On privacy-preserving histograms,” in *Proc. 21st Conf. Uncert. Art. Intell.*, Edinburgh, Scotland, July 2005.

[17] G. Aggarwal, T. Feder, K. Kenthapadi, S. Khuller, R. Panigraphy, D. Thomas, and A. Zhu, “Achieving anonymity via clustering,” in *Proc. Symp. Principles Database Sys.*, Dallas, TX, June 2006.

[18] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, “L-diversity: Privacy beyond k-anonymity,” *ACM Trans. Knowledge Discov. Data*, vol. 1, no. 1, 2007.

[19] A. Sarwate and K. Chaudhuri, “Privacy constraints in regularized convex optimization,” July 2009, arxiv.org e-print 0907.1413v1.

[20] W. Gasarch, “A survey on private information retrieval,” *Bulletin of the EATCS*, vol. 82, pp. 72–107, 2004.

[21] R. Canetti, Y. Lindell, R. Ostrovsky, and A. Sahai, “Universally composable two-party and multi-party secure computation,” in *Proc. 34th Annual ACM Symp. Theory of Computing*, Montreal, Quebec, Canada, may 2002, pp. 494–503.

[22] O. Goldreich, *The Foundations of Cryptography: Basic Tools*. Cambridge, UK: Cambridge University Press, 2001.

[23] Y. Liang, H. V. Poor, and S. Shamai, *Information-theoretic security*. Dundrecht, The Netherlands: Now Publishers, 2009.

[24] L. Sankar, S. R. Rajagopalan, V. Aggarwal, and H. V. Poor, “Utility and privacy in databases: An information-theoretic approach,” 2010, in preparation.

[25] C. E. Shannon, “Coding theorems for a discrete source with a fidelity criterion,” *IRE Nat. Conv. Record*, vol. 7, pp. 325–350, 1959.

[26] A. D. Wyner and J. Ziv, “The rate-distortion function for source coding with side information at the decoder,” *IEEE Trans. Inform. Theory*, vol. 22, no. 1, pp. 1–10, Jan. 1976.

[27] V. Erokhin, “Epsilon-entropy of a discrete random variable,” *Theory of Probability and Applications*, vol. 3, pp. 97–100, 1958.

[28] J. T. Pinkston, “An application of rate-distortion theory to a converse to the coding theorem,” *IEEE Trans. Inform. Theory*, vol. 15, no. 1, pp. 66–71, Jan. 1969.